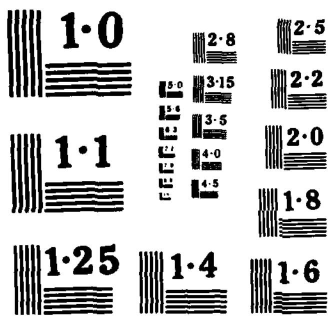
AD-A163 135 EFFECT OF EXAMINEE CERTAINTY ON PROBABILISTIC TEST SCORES AND A COMPARISO..(U) MINNESOTA UNIV MINNEAPOLIS COMPUTERIZED ADAPTIVE TESTING LAB.. 1/1

UNCLASSIFIED D SUHADOLNIK ET AL. JUL 83 RR-83-3 F/G 5/9 NL

END

FILMED

DTIC



NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

F630622 (12)

AD-A163 135

Effect of Examinee Certainty on Probabilistic Test Scores and a Comparison of Scoring Methods for Probabilistic Responses

Debra Suhadolnik
David J. Weiss

Research Report 83-3
July 1983

Computerized Adaptive Testing Laboratory
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

This research was supported by funds from the Air Force Office of Scientific Research, Air Force Human Resources Laboratory, Army Research Institute, and Office of Naval Research, and monitored by the Office of Naval Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 83-3

2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER: AD-A163 135

4. TITLE (and Subtitle): Effect of Examinee Certainty on Probabilistic Test Scores and a Comparison of Scoring Methods for Probabilistic Responses

5. TYPE OF REPORT & PERIOD COVERED: Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Debra Suhadolnik and David J. Weiss

8. CONTRACT OR GRANT NUMBER(s): N00014-79-C-0172

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Minnesota, Minneapolis Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P.E.:61153N Proj.:RR042-04; T.A. RR042-04-01; W.U.: NR150-433

11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington Virginia 22217

12. REPORT DATE: July 1983

13. NUMBER OF PAGES: 30

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This research was supported by funds from the Air Force Office of Scientific Research, The Air Force Human Resources Laboratory, the Army Research Institute, and the Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

- Response formats
- Test item response formats
- Probabilistic responses
- Subjective probabilities
- Reproducing scoring systems
- Confidence-weighting procedures
- Response style variables in probabilistic responses
- Scoring methods for probabilistic responses

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The present study was an attempt to alleviate some of the difficulties inherent in multiple-choice items by having examinees respond to multiple-choice items in a probabilistic manner. Using this format, examinees are able to respond to each alternative and to provide indications of any partial knowledge they may possess concerning the item. The items used in this study were 30 multiple-choice analogy items. Examinees were asked to distribute 100 points among the four alternatives for each item according to how confident

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) they were that each alternative was the correct answer. Each item was scored using five different scoring formulas. Three of these scoring formulas--the spherical, quadratic, and truncated log scoring methods--were reproducing scoring systems. The fourth scoring method used the probability assigned to the correct alternative as the item score, and the fifth used a function of the absolute difference between the correct response vector for the four alternatives and the actual points assigned to each alternative as the item score. Total test scores for all of the scoring methods were obtained by summing individual item scores.

Several studies using probabilistic response methods have shown the effect of a response-style variable, called certainty or risk taking, on scores obtained from probabilistic responses. Results from this study showed a small effect of certainty on the probabilistic scores in terms of the validity of the scores but no effect at all on the factor structure or internal consistency of the scores. Once the effect of certainty on the probabilistic scores had been ruled out, the five scoring formulas were compared in terms of validity, reliability, and factor structure. There were no differences in the validity of the scores from the different methods, but scores obtained from the two scoring formulas that were not reproducing scoring systems were more reliable and had stronger first factors then the scores obtained using the reproducing scoring systems. For practical use, however, the reproducing scoring systems may have an advantage because they maximize examinees' scores when examinees respond honestly, while honest responses will not necessarily maximize an examinee's score with the other two methods. If a reproducing scoring system is used for this reason, the spherical scoring formula is recommended, since it was the most internally consistent and showed the strongest first factor of the reproducing scoring systems.

Contents

Technical Editor: Barbara Leslie Camm

Effect of Examinee Certainty on Probabilistic Test Scores and a Comparison of Scoring Methods for Probabilistic Responses

Psychometricians have searched for many years for a test item format that would allow them to measure individual differences on a variable of interest as accurately and as completely as possible. The multiple-choice item has proven to be a useful tool for assessing kowledge, but there are several problems with this item format. These problems include the possibility of an examinee guessing the correct answer, the lack of information concerning the process used by an examinee to obtain a given answer, and, in general, an inability to accurately determine an examinee's level on a continuous underlying trait based on an observable dichotomous response.

In attempts to remedy these problems and to extract the maximum amount of information from an individual's responses to a set of test items, Lord and Novick (1968, Chap. 14) have identified three important components of interest. These components are

1. The measurement procedure, or the manner in which examinees are instructed to respond to the items.
2. The item scoring formula.
3. The method of weighting each item to form a total score.

In their attempts to find alternatives to the conventional multiple-choice item where the examinee is instructed to choose the one best answer to an item from a number of alternatives, investigators have generally focused on one or two of these components at a time.

The various attempts to improve upon the traditional multiple-choice item can be classified into three broad categories: (1) attempts to improve the multiple-choice item by using an item-weighting formula other than the conventional unit-weighting scheme, (2) variations of the multiple-choice item that attempt to provide more information about an examinee's ability level by asking the examinee to respond to a traditional multiple-choice item in a manner other than simply choosing the one best alternative, and (3) the use of item types which are completely different from the conventional multiple-choice item, such as free-response items. The first category focuses on the third component enumerated by Lord and Novick, the item-weighting formula. The second category focuses on Lord and Novick's first two components--the measurement procedure and item-scoring formulas--while continuing to use a unit-weighting scheme to combine item scores into a total score. The third category focuses primarily on the measurement procedure and, to a lesser extent, on item scoring formulas.

Item-Weighting Formulas

For many years the accepted method of combining item scores to form a test score was simply to sum all of the individual item scores. Since this procedure is equivalent to multiplying each item score by an item weight of 1 and then summing the weighted item scores, the method has been called unit weighting. In attempts to increase the validity and/or the reliability of test scores obtained by summing item scores, many researchers have abandoned unit weighting in favor of various forms of differential weighting of individual items. These methods of differential weighting of items include multiple regression techniques (Wesman & Bennett, 1959), using the validity coefficient of the item as the item weight (Guilford, 1941), weighting items by the reciprocal of the item standard deviation (Terwilliger & Anderson, 1969), a priori item weights (Burt, 1950), and numerous other weighting procedures (Bentler, 1968; Dunnette & Hogatt, 1957; Hendrickson, 1970; Horst, 1936; Wilks, 1938).

In reviewing the substantial literature in this area, Wang and Stanley (1970, p. 664) have concluded that "although differential weighting theoretically promises to provide substantial gains in predictive or construct validity, in practice these gains are often so slight that they do not seem to justify the labor involved in deriving the weights and scoring with them. This is especially true when the component measures are test items ...." Gulliksen (1950) concluded, in concurrence with Wang and Stanley (1970), that differential weighting is not worthwhile when a test contains more than approximately 10 items and when the items are highly correlated. Stanley and Wang (1970), after concluding that differential item weighting is not a fruitful venture for test items, have suggested that the item score be determined by the response made to an item, where the examinee is required to do more than just select the correct alternative for an item. By changing the mode of response and devising item scoring formulas appropriate for these types of responses, the validity and/or reliability of test scores might be increased. An additional gain might be more insight into the process involved in responding to test items.

Variations of the Response Format of Multiple-Choice Items

Several of the earliest attempts at modification of the method of responding to a conventional multiple-choice item were reported by Dressel and Schmid (1953) in an investigation of various item types and scoring formulas. A conventional multiple-choice test and one of four "experimental test forms" were administered to each subject. The items in each of the experimental test forms resembled conventional multiple-choice items in that an item stem and several alternatives were provided, but each experimental test form differed from the conventional multiple-choice format in the following ways:

1. Free-choice format. Examinees were instructed to choose as many of the alternatives provided as necessary to insure that they had chosen the correct alternative. This item format was scored using Equation 1, which yields integer scores that range from -4 to 4 and applies only to five-alternative items:

   $$\text{Item score} = 4C - I \qquad [1]$$

   where C = number of correctly marked alternatives and
   I = number of incorrectly marked alternatives.
2. Degree-of-certainty test. Examinees were instructed to choose the one best answer for an item and then to choose one of four confidence ratings provided to indicate the degree of confidence they had in the answer they had chosen. This item format was scored as shown in Table 1.
3. Multiple-answer format. Each item contained more than one correct alternative, and the examinees were instructed to choose all of the correct alternatives. The score for this format was the number of correct alternatives chosen minus a correction factor for any incorrect alternatives chosen.

Table 1
Scoring System for Degree-of-Certainty Test

Confidence Rating	Item Score: Correct Answer Chosen	Item Score: Incorrect Answer Chosen
Positive	4	-4
Fairly certain	3	-3
Rational guess	2	-2
No defensible basis for choice	1	-1

4. Two-answer format. Each item contained exactly two correct alternatives, and the examinees were instructed to indicate both of the correct alternatives. The item score was simply the number of correct alternatives chosen.

In comparing these five test forms (the conventional multiple-choice format and the four experimental test formats), Dressel and Schmid's (1953) results showed that the experimental test formats containing more than one correct alternative (Formats 3 and 4 above) exhibited greater internal consistency reliability than the other three test forms, but these test formats also took longer to administer than all of the other formats. All of the experimental test formats had higher internal-consistency reliability than the conventional multiple-choice test except for the free-choice format, but the conventional multiple-choice format took less time than any of the experimental test formats. Although the higher reliability coefficients of several of these formats (Formats 2, 3, and 4) might suggest that these formats aid in introducing more ability variance than error variance, the authors warn that the results must be viewed with caution, since there were statistically significant differences between the groups taking each experimental form on the standard multiple-choice test that was administered to all of their subjects; thus, the differences attributed to the effect of test format might be due to systematic ability differences in the groups taking each of the experimental test formats.

Hopkins, Hakstian, and Hopkins (1973) used a confidence weighting procedure similar to the degree-of-certainty test used by Dressel and Schmid (1953) and reported higher split-half reliability coefficients for the confidence weighting format than for a conventional multiple-choice test using the same items. Hopkins et al. (1973) also reported validity coefficients that were correlations between the test scores and a short-answer form of the same test. The validity coefficient for the conventional test (.70) was higher but not significantly different from that of the confidence weighting format (.67).

Coombs (1953) felt that examinees could provide more information about the degree of knowledge they possessed by eliminating the alternatives which they felt were incorrect, rather than by choosing the one correct alternative. Items using this format were scored by assigning one point for each incorrect alternative eliminated and 1 - K points when the correct alternative was eliminated, where K is the number of alternatives provided. This scoring system yields a range of integer item scores from -3 to 3 for a four-alternative multiple-choice item.

In comparing this test format with a conventional multiple-choice test, Coombs, Milholland and Womer (1956) found no differences in validity between the two formats for separate tests of vocabulary, spatial visualization, and driver information. The validity coefficients used were correlations between test scores and criteria such as Stanford-Binet IQ, another test of spatial ability, and subtest scores from the Differential Aptitude Test. For these same content areas, the experimental test format yielded higher reliability estimates than the conventional test, but the differences between the estimates were not statistically significant for any of the content areas. One result in favor of the experimental test format was that the subjects in the experiment felt the experimental format to be fairer than the conventional format.

Another variation upon the conventional multiple-choice item includes a self-scoring method advocated by Gilman and Ferry (1972), which requires examinees to choose among alternatives provided until the correct alternative is chosen. Feedback is given after each choice is made. The item score is simply the number of responses needed to choose the correct alternative; thus, a higher score indicates less knowledge about an item. Kane and Moloney (1974) have warned that although Gilman and Ferry (1972) found an increase in split-half reliability using this technique, the effect of using this method on the reliability of the test depends upon the ability of the distractors to discriminate between examinees of varying levels of ability. An increase in reliability will result when the distractors possess this ability to discriminate among ability levels, but no increase in reliability will occur if this is not the case.

Use of Subjective Probabilities with Multiple-Choice Items

A modification of the traditional multiple-choice item that has generated much research and interest is the use of examinees' subjective probabilities concerning the degree of correctness of each alternative provided for an item as a method of assessing the degree of knowledge or ability possessed by the examinees. By assigning a probability estimate for each alternative to an item, examinees can indicate degrees of partial knowledge they may have concerning each alternative for an item.

To simplify this procedure for examinees, a number of methods have been devised to aid examinees in assigning their subjective probabilities to the alternatives. One method is to ask examinees to directly assign probabilities from 0 to 1.00 to each alternative, with the restriction that the probabilities assigned to all of the alternatives for each item sum to 1.00. Another method instructs examinees to distribute 100 points among the alternatives for each item. The distributed points are then converted to probabilities for scoring purposes by dividing the points assigned to each alternative by 100. Some investigators have used fewer points for distribution (Rippey, 1970) or symbols, such as a certain number of stars, which are to be distributed among the alternatives (deFinetti, 1965), but the concept is the same.

Using these types of measurement procedures (sometimes called probabilistic item formats or probabilistic response formats), an item scoring formula had to be devised so that examinees' expected scores would be maximized only when they responded according to their actual beliefs concerning the correctness of each alternative. Item-scoring formulas which satisfy these conditions are called reproducing scoring systems (RSS). Shuford, Albert, and Massengill (1966) and deFinetti (1965) provide examples of several RSSs. The RSSs presented by these two authors for use with multiple-choice items that have more than two alternatives and only one correct answer are the following:

1. Spherical RSS

$$\text{Item score} = p_c \Big/ \left(\sum_{k=1}^{m} p_k^2 \right)^{1/2} \qquad [2]$$

where p_c = probability assigned to the correct alternative

p_k = probability assigned to alternative k, $k = (1, 2, \ldots, m)$

2. Quadratic RSS

$$\text{Item score} = 2p_c - \sum_{k=1}^{m} (p_k^2) \qquad [3]$$

3. Truncated Logarithmic Scoring System

$$\text{Item score} = \begin{cases} 1 + \log(p_c), & .01 < p_c \leq 1.00 \\ -1, & 0 \leq p_c \leq .01 \end{cases} \qquad [4]$$

or a modification of this scoring function:

$$\text{Item score} = \begin{cases} [2 + \log(p_c)/2], & .01 \leq p_c \leq 1.00 \\ 0, & 0 \leq p_c \leq .01 \end{cases} \qquad [5]$$

The truncated logarithmic scoring system is technically not an RSS, but it does have the properties of an RSS for probabilities between .027 and .973. According to Shuford et al. (1966), when examinees believe that an alternative has a probability of being the correct answer less than or equal to .027, their score will be maximized by assigning a probability of zero to that alternative. Alternatively, when examinees believe that an alternative has a probability greater than or equal to .973, their expected score will be maximized by assigning a probability of 1.00 to that alternative. Shuford et al. (1966) stated that "for extreme values of (p_k), some information about the student's degree-of-belief probabilities is lost, but from the point of view of applications, the loss in accuracy is insignificant" (p. 137). Note also that the truncated logarithmic scoring function is the only one of the scoring formulas that is dependent only upon the probability assigned to the correct alternative.

Total test scores for examinees are obtained for all of the RSSs by simply summing the individual item scores obtained using that particular scoring formula. In addition to the conditions expressed above for an RSS, deFinetti (1965) has stated that the validity of any reproducing scoring system also rests upon the following assumptions:

1. The examinees are capable of assigning numerical values to their subjective probabilities.
2. The examinees are trained in using the response format and understand the scoring system to be used in scoring the items.
3. The examinees are motivated to do their best on the items.

Rippey (1968) reported results from several studies comparing test scores obtained using the spherical RSS and the modification of the truncated logarithmic scoring functions with test scores obtained by summing dichotomous (0,1) item scores to conventional multiple-choice items. In general, he found increases in Hoyt's reliability coefficient using a probabilistic response format with RSSs under limited conditions. The probabilistic test format produced increases in test reliability with undergraduate college students but could not be used with fourth graders and produced no consistent increases in reliability for tests given to high school freshmen or medical students. There were also no consistent tendencies for one or the other of the scoring formulas for the probabilistic response format to produce higher reliability coefficients.

Rippey (1970) compared the reliabilities of five different methods of scoring probabilistic item responses. Three of these methods were RSSs; the fourth was simply the probability assigned to the correct answer, and the fifth was a dichotomous scoring of the probabilistic responses, which resulted in an item score of 1 if the probability assigned to the correct answer was greater than the probability assigned to any other alternative and a score of 0 otherwise. The three RSSs used were the modification of the truncated log scoring function, the spherical RSS, and another RSS called the Euclidean RSS. An item score using the Euclidean RSS is computed using the following equation:

$$\text{Item score} = 1 - \left\{ \left[\sum_{k=1}^{m} \left(p_k - X_k \right) \right]^{1/2} \right\} \Big/ \sqrt{2} \qquad [6]$$

where p_k = probability assigned to alternative $\underline{k}$, $\underline{k}$ = (1, 2, ..., $\underline{N}$), and X_k = criterion group mean probability assigned to alternative $\underline{k}$.

Using Hoyt's reliability coefficient, Rippey found that the test scores obtained by summing the probabilities assigned to the correct answer yielded higher average reliability coefficients (.69) than any of the other scoring methods and that the dichotomous scoring of the probabilistic responses yielded the lowest average reliability of the five methods (.47), although it was not much lower than those of the three RSSs (.49, .50, and .58).

In comparing two RSSs (quadratic and the modification of the truncated logarithmic scoring functions) with conventional multiple-choice test scores, Koehler (1971) found no significant differences between internal consistency reliability coefficients for the test scores obtained using the two RSSs and the test scores from the conventional multiple-choice items. He found evidence of convergent validity for both the probabilistic and conventional item formats and, on the basis of this evidence, suggested the use of conventional tests, since they are "easier to administer, take less testing time, and do not require the training of subjects in the intricacies of the confidence-marking procedures" (p. 302). However, his conclusions must be viewed with caution, since each of his tests consisted of only 10 items.

Extraneous Influences on the Use of Subjective Probabilities with Multiple-Choice Items

Although Koehler's results may not be generalizable due to the small number of items administered in each format, the use of the probabilistic item format has been questioned for other reasons. Hansen (1971), Jacobs (1971), Slakter (1967), Echternacht, Boldt, and Sellman (1972), Koehler (1974), and Pugh and Brunza (1974), along with several others, have investigated the possibility that the increase in reliability demonstrated by probabilistic item formats is due to the effect of a personality variable or response style variable rather than a more accurate assessment of knowledge. This variable has been alternately called risk taking, certainty, confidence, and cautiousness. If it is the effect of this response style variable that leads to increases in reliability for probabilistic responding over conventional multiple-choice items, this effect might also explain the fact that the probabilistic item format has not, in general, led to increases in the validity of these test scores over that of test scores obtained from conventional multiple-choice items.

Studies investigating the influence of these various personality variables have shown mixed results. In studies where conventional multiple-choice item scores and probabilistic item scores were obtained (Koehler, 1974; Echternacht, Sellman, Boldt, & Young, 1971), the correlations between the two types of scores have been consistently high (.71 to .83 for the Koehler (1974) study and .89 to .99 for the Echternacht et al. (1971) study). This suggests that a large proportion of the variation in the probabilistic test scores can be accounted for by the conventional test scores. The question being posed, though, is whether the variation in the probabilistic test scores that cannot be accounted for by the conventional test scores is reliable variance due to increased accuracy of assessment of knowledge or due to personality or response style variables.

To determine the influence of these personality factors, Koehler (1974) embedded seven nonsense items in a 40-item vocabulary test and told examinees that they were not to guess the answers to any items on the test. The nonsense items were items with no correct alternatives. From responses to these nonsense items he calculated two confidence measures:

C_1 = proportion of nonsense items attempted under do-not-guess instructions,

and

$$C_2 = \sum_{j=1}^{n} \sum_{i=1}^{m} \left(p_{ij} - \frac{1}{m}\right)^2 \Big/ \left(1 - \frac{1}{m}\right) \qquad [7]$$

where $\underline{m}$ = number of alternatives,

$\underline{n}$ = number of nonsense items, and

p_{ij} = probability assigned to alternative $\underline{i}$ on item $\underline{j}$.

Since the nonsense items had no correct alternatives, an examinee's responses to these items were a pure measure of a response style or personality variable (confidence) that was influencing that examinee's responses. Responses to these items were not due to any knowledge the examinee possessed, since there were no correct answers to those items. The greater the deviation of these indices from 0, the higher the level of confidence exhibited by the examinee.

Koehler found that both of these confidence indices were significantly negatively correlated with three probabilistic test scores (spherical, quadratic, and the modification of the truncated logarithmic scoring functions), but not significantly correlated with the number-correct scores from the same items. The number-correct scores also yielded a higher internal consistency reliability coefficient than the three probabilistic scores (.85 versus .82, .80, and .74). On the basis of these results, Koehler did not recommend the use of probabilistic response formats, since "it would appear ... that confidence responding methods produce variability in scores that cannot be attributed to knowledge of subject matter" (p. 4).

Hansen (1971) obtained probabilistic test scores and scores on independent measures of personality factors such as risk taking and test anxiety. He developed a measure of certainty in responding to probabilistic response formats which is essentially the average absolute deviation of a response vector to an item from a response vector assigning equal probabilities to all alternatives. Hansen's study showed that this certainty index was related to risk taking as measured by the Kogan and Wallach Choice Dilemmas Questionnaire and authoritarianism as measured by a version of the F-scale, developed by Christie, Havel, and Seidenberg (1958). However, the certainty index did not correlate significantly with scores on a test anxiety questionnaire or scores on the Gough-Sanford Rigidity Scale.

These results provide more information concerning the nature of the response style, but there are problems with Hansen's (1971) certainty index, which he attempts to alleviate but does not. The major problem with this index is that it is not a pure measure of certainty. This certainty measure is confounded by an examinee's knowledge concerning an item. Hansen attempted to partial out examinees' knowledge by using their test scores as a predictor in a regression equation to obtain predicted certainty scores. These predicted certainty scores were then subtracted from the observed certainty scores to obtain a certainty measure free of the influence of examinee knowledge.

Although the rationale is sound, Hansen did not accomplish what he set out to do. The test score he used as a predictor was not a pure or even relatively pure measure of knowledge. The test scores were probabilistic test scores computed from the spherical RSS. This scoring system results in scores that represent a confounding of certainty and knowledge. Therefore, by partialling these probabilistic test scores from the certainty index, it is unclear exactly what the residual certainty index represents, since both knowledge and some certainty have been partialled out. Hansen's results were then based upon the relationship of various personality variables with a certainty index confounded with knowledge, and the relationship of these same personality variables with a residual certainty index whose composition is somewhat ambiguous. Hansen's results might best be viewed with caution.

Pugh and Brunza (1974) conducted a study similar to that of Hansen (1971), except that they used a 24-item vocabulary test and scored it using the probability assigned to the correct answer as the item score. They also obtained scores on an independent nonprobabilistically scored vocabulary test, and measures of risk taking, degree of external control, and cautiousness. They followed Hansen's regression procedure to obtain a certainty measure free of the confounding effects of knowledge and were more successful than Hansen. They used the independent vocabulary tes' score as a predictor of the same certainty index that Hansen used and then calculated a residual certainty index by subtracting the predicted certainty score from the observed certainty score. Since the independent vocabulary test was a relatively pure measure of knowledge, partialling its effect from the observed certainty index resulted in a residual certainty index that (1) was a measure of the certainty displayed in responding to multiple-choice items in a probabilistic fashion and (2) was not related to knowledge possessed by examinees concerning the items.

Pugh and Brunza (1974) reported that this residual certainty measure was not very reliable (.32 internal consistency reliability) and that it correlated significantly with risk-taking scores obtained from the Kogan and Wallach Choice Dilemmas Questionnaire but not with the measures of cautiousness and external control they had obtained. Although this evidence of the influence of variables other than knowledge on probabilistic test scores might serve as a deterrent to the use of these scoring systems, Pugh and Brunza noted that "there is no evidence in either study [Pugh & Brunza, 1974, or Hansen, 1971] that these factors are more operative than in traditional tests" (p. 6).

Echternacht et al. (1971) scored answer sheets of daily quizzes obtained from two Air Force training courses using a truncated logarithmic scoring function and number correct. They found that using the number-correct score, the shift of the trainees, and a number of personality variables such as test anxiety, risk taking, and rigidity as predictors of the probabilistic test scores did not account for significantly more of the variation in the probabilistic test scores than was accounted for when using only number-correct scores and shift of the trainees as predictors. This is evidence that the personality variables did not operate to a greater extent in a probabilistic testing situation than in a conventional multiple-choice testing situation.

Thus, these studies show some relationship of probabilistic test scores to personality variables (primarily risk-taking tendencies); but they also show that these influences do not seem to be greater in probabilistic testing situations than in conventional testing situations.

Use of Alternate Item Types

The research reviewed above relied on the multiple-choice item type and varied the method of responding to that type of item; however, some researchers have advocated the use of entirely different item types, such as free-response items, to aid in the assessment of partial knowledge. Some of these alternate item types avoid many of the problems inherent in multiple-choice items but are subject to problems of their own. For example, the free-response item type avoids the problem of random guessing among a number of alternatives and has the potential to provide a large amount of information concerning what the examinee does or does not know, but it is also more time-consuming to administer and score, and may cover much less material than is possible with a multiple-choice format. Consequently, if there are any time constraints on testing, fewer items can be administered. Practical problems with scoring many of these alternate item types have prevented widespread use of several of them.

Purpose

Although comparisons of the psychometric properties of multiple-choice items with several alternate item types are planned, the present research focused on comparisons of the probabilistic response formats. This study has attempted to answer the following questions:

1. Does a personality variable such as certainty affect probabilistic test scores on an ability test to a greater degree than it affects conventional test scores on the same ability test?

2. If the effect of a personality variable can be discounted, what types of scoring systems are best for multiple-choice items on an ability test requiring probabilistic responses?

Method

Test Items

Thirty multiple-choice analogy items were chosen from a pool of items obtained from Educational Testing Service (ETS) containing former SCAT and STEP items. Each item consisted of an item stem and four alternatives. The pool of items had been parameterized by ETS on groups of high school students using the computer program LOGIST (Wood, Wingersky, & Lord, 1976) with a three-parameter logistic model, resulting in item response theory discrimination, difficulty, and guessing parameters calculated from large numbers of examinees for each item. The 30 items were chosen from a pool of approximately 300 analogy items to represent a uniform range of discrimination and difficulty parameters. The parameters for the chosen items are in Appendix Table A. The item discrimination parameters ranged from approximately $a = .6$ to $a = 1.4$, with a mean of .975 and a standard deviation of .244, while the difficulty parameters ranged from approximately $b = -.5$ to $b = 2.5$, with a mean of .961 and a standard deviation of .887. The range of difficulty parameters was not chosen to be symmetric about zero because the available examinees constituted a more select group than the group whose responses were used to parameterize the items. The guessing parameters for these items ranged from $c = .09$ to $c = .38$, with a mean of .20 and a standard deviation of .06.

Test Administration

The 30 multiple-choice analogy items chosen were then administered to 299 psychology and biology undergraduate students at the University of Minnesota during the 1979-1980 academic year. Students received two points toward their course grade (either introductory psychology or biology) for their participation. Items were administered by computer to permit checking of responses to be sure that item response instructions were carefully followed.

The examinees were instructed to respond to each item by assigning a probability to each of the four alternatives. This probability was to correspond to the examinee's belief in the correctness of each alternative, with the additional restriction that the probabilities assigned to all of the alternatives for an item sum to one. Specifically, for each item, the examinees were asked to distribute 100 points among the four alternatives provided for each item according to their belief as to whether or not the alternative was the correct alternative for that item. The total number of points assigned to all of the alternatives for an item had to equal 100. Since the tests were computer administered, item responses were summed immediately to ensure that the responses to the alternatives did indeed sum to 100 (sums of 99 and 101 were also considered valid to allow for rounding). The points assigned to each alternative were then converted into probabilities by dividing the response to each alternative by 100.

To insure that the examinees understood both how to use the computer and how to respond to the multiple-choice items in a probabilistic fashion, a detailed set of instructions preceded each test (see Appendix Table B). If an examinee responded incorrectly to an instruction, the computer would display an appropriate error message on the CRT screen and the examinee would have to respond correctly before proceeding to the next screen. If an examinee again responded inappropriately to an instruction, a test proctor was called by the computer to provide additional help to the examinee in understanding the instructions. Several examples and explanations of methods of responding to probabilistic items were provided. Examinees, with few exceptions, did not have any difficulty understanding how to respond to the items. If, in responding to an item, an examinee's responses did not sum to 99, 100, or 101, the examinee was immediately asked to reenter his/her responses until an appropriate sum was entered.

Item Scoring

The item responses obtained from these 299 examinees were then scored using five different scoring formulas to determine which of these scoring formulas yielded the most reliable and valid scores. The five different scoring formulas used were:

1. The probability assigned to the correct alternative by the examinee (PACA) was used as the item score. This scoring formula yields scores that range from 0 to 1.00.
2. The second type of item score (AIKEN) was computed from a variation of a scoring formula developed by Aiken (1970), which is a function of the absolute difference between the correct response vector for an item and the obtained response vector:

$$\text{Item score} = 1 - \frac{D}{D_{max}} \qquad [8]$$

$$\text{where } D = \sum_{i=1}^{m} \left| p_{ai} - p_{ei} \right| \qquad [9]$$

$\underline{m}$ = number of alternatives,
p_{ai} = probability assigned to the alternative by the examinee;
p_{ei} = expected probability for alternative; and
D_{max} = maximum value of D, which was 2.00 for all of these items.

Each correct response vector would contain three 0's and one 1, while the obtained response vector would contain four probabilities that sum to 1.00. For example, for an item where the second alternative was the correct alternative, the correct response vector would be 0, 1.00, 0, 0. A response vector that might have been obtained for this item is .20, .60, .20, 0. For this obtained response vector the item score would be computed as follows:

$$\text{Item score} = 1 - \left[\frac{|0-.20| + |1.00-.60| + |0-.20| + |0-0|}{2.00}\right]$$

$$= 1 - \frac{.80}{2.00} = .60 \qquad [10]$$

This scoring formula also yields scores that range from 0 to 1.00.

3. The quadratic RSS (QUAD), is defined by Equation 3. This scoring formula yields scores that range from -1.00 to 1.00.
4. The spherical RSS (SPHER) is defined in Equation 2. This scoring formula yields scores that range from 0 to 1.00.
5. A modification of the truncated logarithmic scoring function (TLOG). This scoring formula is a good approximation to the logarithmic Rss. It is a very good approximation throughout most of the possible score range, and is defined by Equation 5. This scoring formula yields scores from 0 to 1.00. The actual formula used here to obtain scores via a truncated logarithmic scoring function utilizes a scaling factor of 5 rather than the usual scaling factor of 1 or 2. It was necessary to increase this scaling factor to maintain a logical progression of scores, since the probability assigned to the correct answer for some items was as low as .01. Since the log of .01 is -4.6052, the scaling factor had to be a 5 (actually only some number slightly higher than 4.6052) in order that the scores progress in an orderly fashion from 0 to 1.00 according to the probability assigned to the correct answer. This alleviated the problem of assigning negative scores to examinees who had assigned very small probabilities to the correct answer while assigning a score of 0 (a higher score) to examinees who had assigned a zero probability to the correct answer. The actual TLOG scoring formula used is Equation 11.

$$\text{Item score} = \begin{Bmatrix} \dfrac{5 + \log(p_c)}{5}, & .01 \leq p_c \leq 1.00 \\ \\ 0, & 0 \leq p_c \leq .01 \end{Bmatrix} \qquad [11]$$

Total test scores for all of the scoring methods were obtained by summing all 30 item scores for each of the 30 items.

Determining the Effect of Certainty

To determine the effect of an examinee's certainty or propensity to take risks when responding to probabilistic items, Hansen's (1971) certainty index was computed for each examinee using the following formula:

$$C_T = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{m_j}{2(m_j - 1)} \right) \sum_{i=1}^{m} \left| \frac{1}{m_j} - p_{ij} \right| \qquad [12]$$

where

C_T = certainty index,

$\underline{n}$ = number of items in test,

m_j = number of alternatives for item $\underline{j}$, and

p_{ij} = probability assigned to alternative $\underline{i}$ of item $\underline{j}$.

This certainty index is a function of the absolute difference between the probabilities assigned to the four alternatives and .25, averaged over items. Since the probabilities assigned to each alternative are dependent upon both an examinee's knowledge and his/her level of certainty, this certainty index is not a "pure" measure of certainty, but is confounded with knowledge about the item.

To determine the effect of this response style variable, it was first necessary to obtain a "pure" measure of certainty. This relatively pure measure of certainty was obtained by scoring the probabilistic responses dichotomously and then partialling the effect of this knowledge variable out of the certainty indices. A dichotomous test score was obtained from the probabilistic responses by making the assumption that under conventional "choose-the-correct-answer" instructions, examinees would choose the alternative to which they assigned the highest probability under the probabilistic instructions. Thus, for each item, the alternative assigned the highest probability by the examinee was chosen as the alternative the examinee would have chosen under traditional multiple-choice instructions. A score of 1 was assigned if that alternative was the correct answer and a score of 0 was assigned otherwise. When more than one alternative was assigned the highest probability, one of those alternatives was randomly chosen as the alternative the examinee would have chosen. This procedure attempted to simulate the decision-making process of an examinee in choosing a correct answer to an item.

This dichotomous test score was used in a regression equation to predict the certainty index. The predicted certainty index was then subtracted from the actual certainty index to obtain a residual certainty index. This residual certainty index constituted a "pure" measure of certainty. This pure certainty index was partialled out of the probabilistic test scores using the same method as that used to partial the dichotomous test scores out of the original certainty index. The pure certainty index was also used to predict the probabilistic test score. The predicted probabilistic test score was then subtracted from the probabilistic test score to obtain a residual probabilistic test score that was unassociated with the pure certainty index.

As a result of these partialling operations, the following measures were available for each of the five scoring methods:

1. <u>Probabilistic test score</u>. This score represents a confounding of knowledge and certainty.
2. <u>Dichotomous test score</u>. This score represents a pure knowledge index

and is the dichotomous scoring of the probabilistic responses.

3. Residual score. This score is the probabilistic test score with the pure certainty index partialled out, and thus represents the pure knowledge component of the probabilistic scores.
4. Certainty index. This measure represents a confounding of knowledge and certainty.
5. Residual certainty index. This measure is the certainty index with the pure knowledge index (the dichotomous test score) partialled out and thus represents a pure certainty index.

Evaluative Criteria

Reliability and validity coefficients were computed for both the probabilistic and the residual test scores. The reliability coefficients were internal consistency reliability coefficients calculated using coefficient alpha. The validity coefficients were the correlations between test score and reported grade-point average. For each of the five scoring methods used, the validity and reliability of the residual scores was compared with that of the original probabilistic test scores. If there was any difference between the validities and the reliabilities of the probabilistic and the residual scores, they could be attributed to the effect of certainty in responding, since the only difference between the two scores was that the effect of certainty had been removed from the residual scores.

Factor analyses of the item scores (both probabilistic and residual) for each of the five scoring formulas were performed using a principal axis factor extraction method. The number of factors extracted for each of the scoring formulas was determined through parallel analyses (Horn, 1965) performed separately for each scoring formula, using randomly generated data with the same numbers of items and examinees as the real data and with item difficulties (proportion correct) equated with the real data. Coefficients of congruence and correlations between factor loadings for each of the five scoring formulas were computed.

Results

Score Intercorrelations

Correlations between probabilistic test scores, residual test scores, dichotomous scores, the certainty index, and the residual certainty index for each of the scoring formulas are presented in Table 1. Since the AIKEN scoring formula resulted in item scores and correlations that were identical to that of the PACA scoring formula, only the PACA results are reported.

As expected, due to the partialling procedure, the correlation between the residual certainty index and the dichotomous score, and the correlation between the residual certainty index and the residual score, were both zero for all scoring methods. The correlation between the original certainty index and the dichotomous score (.71), and the correlation between the original certainty index and the residual certainty index (.71), were exactly the same for all four scoring formulas. This is due to the fact that the three indices--the original certainty index, the residual certainty index, and the dichotomous score--do not

Table 1
Intercorrelations of Scores for Multiple-Choice Items with a Probabilistic Response Format Scored by Four Scoring Methods

Scoring Method and Score	Probabilistic	Dichotomous	Certainty	Residual Certainty	Residual Score
Quadratic RSS (lower triangle) and Spherical RSS (upper triangle)					
Probabilistic	---	.94**	.64**	-.04	1.00**
Dichotomous	.91**	---	.71**	.00	.94**
Certainty	.56**	.71**	---	.71**	.67**
Residual Certainty	-.12*	.00	.71**	---	-.00
Residual Score	.99**	.92**	.65**	.00	---
Truncated Log RSS (lower triangle) and PACA (upper triangle)					
Probabilistic	---	.93**	.83**	.24**	.97**
Dichotomous	.85**	---	.71**	.00	.96**
Certainty	.43**	.71**	---	.71**	.68**
Residual Certainty	-.25**	.00	.71**	---	-.00
Residual Score	.97**	.88**	.62**	.00	---

*$p < .05$
**$p < .01$

change with the particular scoring formula used; they are constant for each individual across scoring methods. These two significant correlations, along with the significant correlations exhibited for each of the scoring formulas between the certainty index and the residual score (.65, .67, .62, and .68 for QUAD, SPHER, TLOG, and PACA, respectively), show that the original certainty index is indeed related to both "knowledge" as measured by traditional multiple-choice tests (the dichotomous scores) and "certainty" unconfounded with "knowledge" (the residual certainty index).

The correlations between the probabilistic test scores and the dichotomous test scores were .91, .94, .85, and .93 for the QUAD, SPHER, TLOG, and PACA scoring methods, respectively. Using approximate significance tests for correlations obtained from dependent samples (Johnson & Jackson, 1959, pp. 352-358), all of the pairwise comparisons among these correlations were significantly different from each other at the .05 level of significance. Practically, the only correlation of these four that appears different from the others is that of TLOG (.85 as opposed to .91, .94, and .93 for the other scoring methods). Squaring these four correlations yields the proportion of variance in the probabilistic test scores accounted for by the dichotomous test scores. The squared correlations are .83, .88, .72, and .86 for the QUAD, SPHER, TLOG, and PACA scoring procedures.

The correlations between the residual certainty index (the "pure" certainty measure) and the probabilistic test scores were -.12, -.04, -.25, and .24 for the QUAD, SPHER, TLOG, and PACA scoring formulas, respectively. The correlations for the QUAD and SPHER scoring formulas were not significantly different from zero at the .01 level of significance and thus do not account for significant amounts of the variance of the probabilistic test scores. Squaring the correlations that are significantly different from zero results in squared correlations of .06 for both the TLOG and PACA scoring formulas. Thus, certainty as measured by the residual certainty index accounts for no more than 6% of the variance of any of the probabilistic test scores.

The correlations in Table 1 between the probabilistic test scores and the residual scores are very high for all four scoring formulas (.99, 1.00, .97, and .97, for QUAD, SPHER, TLOG, and PACA, respectively). These correlations are highest (.99 and 1.00) for the QUAD and SPHER scoring formulas, whose correlations between the probabilistic test score and residual certainty index were not significantly different from zero (-.12 and -.04); these correlations squared (.98 and 1.00) show that almost all of the variance in the QUAD probabilistic test scores, and all of the variance of the SPHER probabilistic test scores, is accounted for by the residual scores (representing "knowledge" concerning the items).

The correlations between the dichotomous test scores and the residual scores are high and significantly different from zero for all of the scoring formulas (.92, .94, .88, and .96 for QUAD, SPHER, TLOG, AND PACA scoring formulas, respectively). This result is expected, since both the residual scores and the dichotomous scores are relatively pure measures of knowledge.

It was also expected that the correlations between the original certainty index and the probabilistic test scores for the various scoring methods would be greater than the correlations between this certainty index and the dichotomous scores, since the probabilistic test scores and the original certainty index both represent a confounding of certainty and knowledge, while the dichotomous scores are a measure of knowledge less confounded by certainty. This occurred only for the PACA scoring method, which was the only scoring method that was not an RSS. The correlation between the certainty index and probabilistic test score was significantly greater than the correlation between the dichotomous score and the certainty index (.83 vs.71) for the PACA scoring formula, and was significantly less (using the dependent samples test of significance for correlations and a .05 level of significance) than .71 (.56, .64 and .43) for the other three scoring formulas.

Validity and Reliability

Table 2 shows the validity and internal consistency reliability coefficients for the probabilistic test scores obtained from the various methods of scoring the multiple-choice items with a probabilistic response format. The validity coefficients were all significantly different from zero but were not significantly different from each other, using a dependent samples test of significance for correlation coefficients (Johnson & Jackson, 1959, pp. 352-358) and maintaining the experimentwise error at a .01 alpha level.

The reliability coefficients were all significantly different from zero and significantly different from each other (using the Pitman procedure described in Feldt, 1980, for testing the significance of differences between coefficient alpha for dependent samples using a .01 significance level). The PACA scoring method yielded the highest internal consistency reliability (.91) followed by SPHER (.88), QUAD (.87), and TLOG (.84).

Table 2
Validity Correlations of Test Scores with Reported GPA and Alpha Internal Consistency Reliability Coefficients for Multiple-Choice Items with a Probabilistic Response Format (N=299)

Scoring Method	Validity r	Validity p*	Reliability α	Reliability p*
Unpartialled Scores				
Quadratic RSS	.18	<.001	.87	<.001
Spherical RSS	.18	<.001	.88	<.001
Truncated Log RSS	.18	<.001	.84	<.001
PACA	.17	<.001	.91	<.001
Residual Scores				
Quadratic RSS	.13	.011	.87	<.001
Spherical RSS	.13	.011	.88	<.001
Truncated Log RSS	.14	.006	.84	<.001
PACA	.12	.017	.91	<.001

*Probability of rejecting null hypothesis of no significant difference from zero.

Validity and internal consistency reliability coefficients for the residual scores are also shown in Table 2. The reliability coefficients for the residual scores are exactly the same as the reliability coefficients for the probabilistic test scores. The validity coefficients for the residual scores were all significantly different from zero but not from each other (.01 significance level), and these validity coefficients were significantly lower ($p \leq .05$) for the residual scores than for the unpartialled probabilistic test scores (.18 vs. .13 for QUAD, .18 vs. .13 for SPHER, .18 vs. .14 for TLOG, and .17 vs. .12 for PACA). This decrease in the magnitude of the validity coefficients of the residual scores is not due to a restriction in range problem, since the range of scores for the probabilistic test scores was very similar to that of the residual scores, as is shown in Table 3.

Table 3
Range of Scores for Probabilistic and Residual Test Scores

Scoring Method	Probabilistic	Residual
Quadratic	27.21	27.30
Spherical	16.57	16.56
Truncated Log	13.14	12.74
PACA	20.69	20.10

Factor Analysis of Probabilistic Test Scores

Factor analyses of the unpartialled probabilistic and residual test scores yielded virtually identical results; therefore, only the results of the factor analyses of the probabilistic test scores are reported here.

Figures 1a to 1d show the results of the parallel analyses performed for each of the scoring methods (numerical data are in Appendix Table C). The eigenvalues obtained from the principal axes factor analysis of the random data were all low; as expected, no factor accounted for significantly more variation in the items than any other factor. In comparing the eigenvalues of the actual data with those from the random data, it is clear that one strong factor is present for all of the scoring methods. A second factor also appears for each of the scoring methods with eigenvalues greater than that of the second factor for the random data, but the eigenvalue for the second factors of the random and actual data are so close that the second factor (and third factor for TLOG) for the actual data can be considered to be the same strength as a random factor. On the basis of these results, one-factor principal axis factor solutions were obtained for each of the scoring methods and are shown in Table 4.

The factor loadings in Table 4 are positive and fairly high for all items and all scoring formulas, indicating a global factor for each of the scoring methods. The magnitudes of the eigenvalues show that this factor accounted for more of the variance of the item responses for the PACA scoring formula (26%) than for any of the other scoring formulas (19.9%, 20.9%, and 17.4% for the QUAD, SPHER, and TLOG scoring formulas).

The correlations between factor loadings across the 30 items for the various scoring methods are presented in the lower left triangle of Table 5, while coefficients of congruence are reported in the upper right triangle of Table 5. The coefficients of congruence are at the maximum of 1.00 for all of the pairs of factor loadings and the correlations among all of the factor loadings are very high, except for the correlation between the factor loadings for the PACA and TLOG scoring methods, which was only .80. The fact that all of the coefficients of congruence are equal to the maximum value for this index is due to the dependence of this index upon the magnitude and sign of the factor loadings. Gorsuch (1974, p. 254) notes that this index will be high for factors whose loadings are approximately the same size even if the pattern of loadings for the two factors is not the same.

Discussion and Conclusions

The Influence of Certainty

The evidence concerning the effect of examinee certainty on probabilistic test scores suggests that certainty as a response style variable has a small, almost negligible effect, on the probabilistic test scores obtained in this study. The reliability coefficients for the five scoring methods were exactly the same for the probabilistic and residual test scores, indicating that the certainty variable was not contributing reliable variance to the probabilistic test scores and was artifically increasing the reliability coefficients. The factor structures of the probabilistic test scores and the residual test scores

Figure 1
Eigenvalues from Parallel Analysis of *Random Data*
and Actual Data for QUAD, SPHER, PACA, and TLOG Scoring Methods

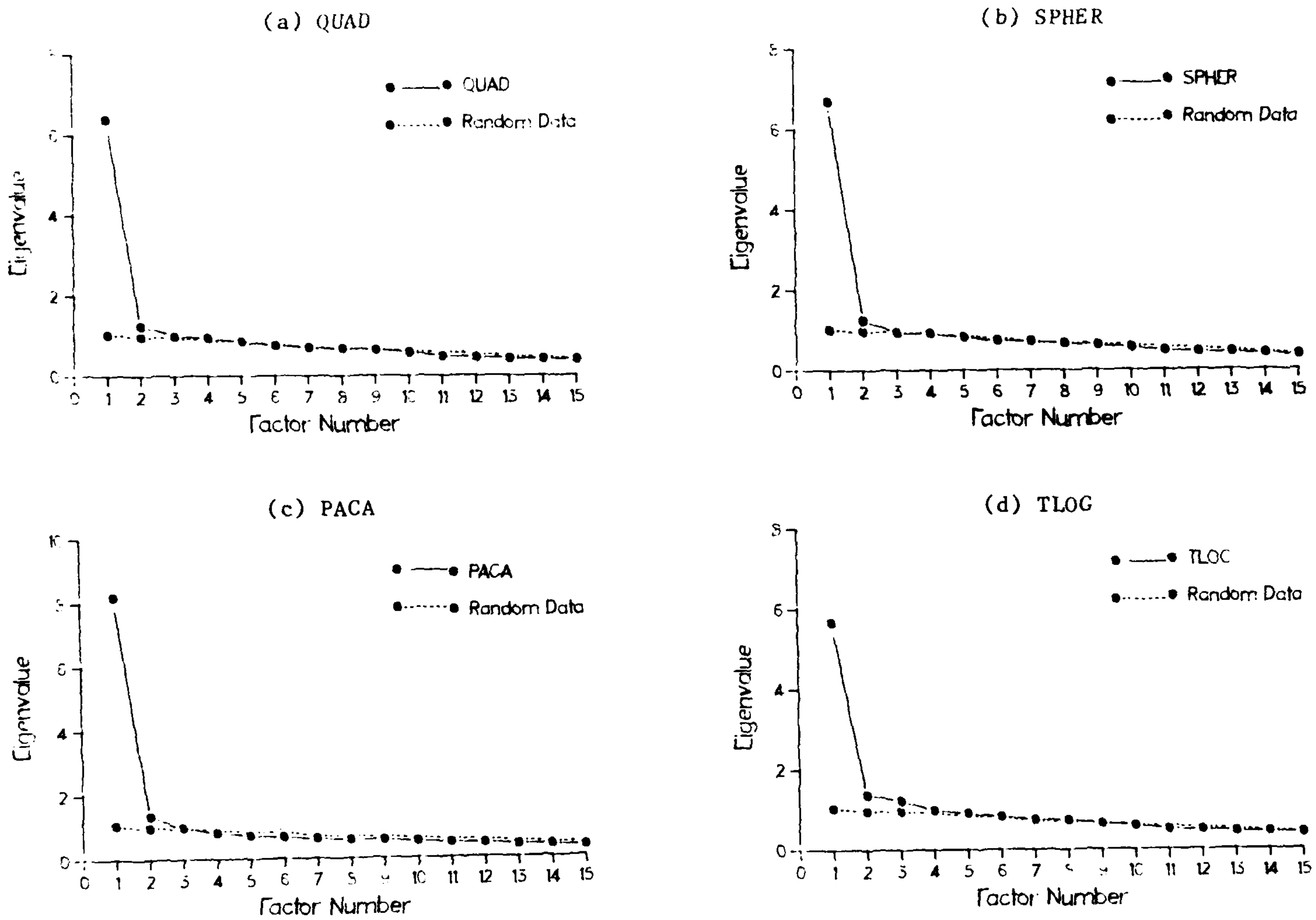


Table 4
Factor Loadings on the First Factor
for Multiple-Choice Items with a
Probabilistic Response Format

Item Number	Scoring Method QUAD	SPHER	PACA	TLOG
1	.418	.433	.382	.490
2	.446	.458	.412	.493
3	.439	.456	.409	.476
4	.439	.435	.358	.526
5	.233	.264	.165	.347
6	.429	.443	.396	.528
7	.332	.358	.316	.412
8	.424	.428	.413	.505
9	.324	.354	.259	.469
10	.426	.414	.391	.500
11	.383	.377	.355	.445
12	.538	.529	.509	.585
13	.513	.513	.519	.566
14	.444	.441	.422	.483
15	.368	.384	.341	.414
16	.465	.512	.469	.543
17	.543	.537	.487	.586
18	.505	.484	.546	.509
19	.316	.338	.244	.445
20	.485	.490	.492	.502
21	.552	.552	.491	.597
22	.544	.571	.518	.624
23	.498	.503	.463	.527
24	.472	.505	.394	.553
25	.400	.422	.380	.466
26	.437	.466	.406	.517
27	.514	.505	.508	.520
28	.524	.515	.473	.571
29	.406	.423	.349	.488
30	.387	.453	.370	.514
Eigenvalue	5.98	6.27	5.22	7.81

Table 5
Correlations (Lower Triangle) and Coefficients
of Congruence (Upper Triangle) Between
Factor Loadings Obtained for Four Scoring Methods

Scoring Method	QUAD	SPHER	TLOG	PACA
QUAD	-	1.00	1.00	1.00
SPHER	.97	-	1.00	1.00
TLOG	.95	.92	-	1.00
PACA	.90	.93	.80	-

were also identical. The factor structure and internal consistency reliability data (which are both based upon the interitem correlations for each scoring method), indicate no effect of the certainty variable on probabilistic test scores above and beyond the effect on the residual test scores (i.e., the probabilistic test scores with the "pure" certainty index partialled out). This lack of effect is demonstrated by the extremely high correlations between the scores derived assuming conventional multiple-choice instructions (the dichotomous score) and the probabilistic test scores for all of the scoring methods studied, and by the extremely low correlations between the "pure" certainty index (the residual certainty index) and the probabilistic test scores for each scoring method. Since the dichotomous test scores simulate testing conditions under conventional multiple-choice instructions to choose the one correct answer, these high correlations suggest that the greatest portion of the variability in the probabilistic test scores for all of the scoring formulas is not different from that present in scores obtained with traditional multiple-choice tests.

The validity coefficients did show an effect of the certainty index on the probabilistic test scores. The significant decrease in the validity coefficients which occurs when the "pure" certainty index is partialled from the probabilistic test scores is evidence of some effect of the certainty variable on the probabilistic test scores. However, even though the decrease was significant for all of the scoring formulas, the practical difference was small. The validity coefficients of the probabilistic test scores were all low initially, since the reported GPA criterion is a complex variable not easily predicted by a single factor of analogical reasoning. Although reported GPA might not have been a true reflection of actual GPA (although Thompson and Weiss, 1980, data show a correlation of .59 between the two), this invalidity should not have affected the comparisons made in this study. Additional research utilizing different criterion measures is recommended to further investigate the generality of the results obtained here.

Other than the small effect of the certainty variable on the validity coefficients for each of the scoring formulas, there appears to be no effect of the certainty variable on the probabilistic test scores. However, since not all of the variance in the probabilistic test scores can be accounted for by the "pure" knowledge and certainty indices, there may be some other response style variable that exerts an influence upon the probabilistic test scores. This influence would have to be extremely small, though, since the knowledge and certainty indices accounted for 88%, 84%, 78%, and 92% of the variance in the scores obtained from the spherical, quadratic, truncated log, and PACA scoring formulas, respectively.

Choice among Scoring Methods

The choice among the five scoring methods must be made on the basis of validity coefficients, the reliability coefficients, and the factor analysis results. Since there were no significant differences between any of the validity coefficients, these coefficients do not provide support for any one scoring method. In terms of the reliability coefficients, the PACA (and its equivalent AIKEN) scoring formula yielded scores having the highest reliability coefficients of all of the scoring methods.

The dependence of both the internal consistency reliability coefficient and the one-factor solution on the interitem correlation suggests that scores from the scoring formulas with the highest reliability coefficients would also have the strongest first factors, and this is exactly what occurred in this study. Hypothesizing that the factor extracted represents verbal ability, it is desirable that this factor account for as large a proportion of each item's variance as possible. The factor contribution of this first factor was greater for the two scoring methods that are not reproducing scoring systems (PACA and AIKEN) than for the three scoring methods that are reproducing scoring systems.

On the basis of these results, either the PACA or Aiken scoring methods can be recommended for use with multiple-choice items with a probabilistic response format. Since PACA is the simplest of the two methods, it might be the preferable scoring method.

Conclusions

Test scores obtained from the five methods of scoring multiple-choice items with a probabilistic response format do not appear to be affected by the response style or personality variable of examinee certainty to a greater degree than scores obtained under traditional multiple-choice instructions. The scoring method used does not affect the validity of the test scores but does appear to affect the internal consistency of the scores. Test scores obtained using the PACA scoring method were more reliable, simpler to compute, and as valid as those obtained from the other scoring methods; therefore, use of the PACA scoring method is recommended for these types of items.

As a note of caution, however, one of the three reproducing scoring systems might have a practical advantage over either the PACA or AIKEN scoring formulas. In a situation where examinees were aware of the scoring formula to be used and where the scores were of some importance to the examinee (as for a classroom grade or selection procedure), the examinees could optimize their test score using the reproducing scoring systems only by responding according to their actual beliefs in the correctness of each alternative, while their total scores could be maximized with the PACA scoring formula by assigning the maximum probability of 1.00 to the one alternative they thought was the correct one. If examinees were expected to utilize this strategy, one of the reproducing scoring systems would be better to use with multiple-choice items with a probabilistic response format. Test scores obtained from the spherical reproducing scoring system were more reliable, as valid, and showed a stronger first factor than scores from the other reproducing scoring systems. Thus, if the practical situation requires use of a reproducing scoring system, the spherical RSS should be used.

References

Aiken, L. R. Scoring for partial knowledge on the generalized rearrangement item. Educational and Psychological Measurement, 1970, 30, 87-94.

Bentler, P. M. Alpha-maximized factor analysis: Its relation to alpha and canonical factor analysis. Psychometrika, 1968, 33, 335-345.

Christie, R., Havel, J., & Seidenberg, B. Is the F scale irreversible? Journal of Abnormal and Social Psychology, 1958, 56, 143-159.

Coombs, C. H. On the use of objective examinations. Educational and Psychological Measurement, 1953, 13, 308-310.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

Dressel, P. L., & Schmid, J. Some modifications of the multiple-choice item. Educational and Psychological Measurement, 1953, 13, 574-595.

Dunnette, M. D., & Hogatt, A. C. Deriving a composite score from several measures of the same attribute. Educational and Psychological Measurement, 1957, 17, 423-434.

Echternacht, G. J., Sellman, W. S., Boldt, R. F., & Young, J. D. An evaluation of the feasibility of confidence testing as a diagnostic aid in technical training (RB-71-51). Princeton NJ: Educational Testing Service, October 1971.

Echternacht, G. J., Boldt, R. F., & Sellman, W. S. Personality influences on confidence test scores. Journal of Educational Measurement, 1972, 9, 235-241.

Feldt L. S. A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 1980, 45, 99-105.

Gilman, D. A., & Ferry, P. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 1972, 9, 205-207.

Gorsuch, R. L. Factor analysis. Philadelphia: W. B. Saunders Company, 1974.

Guilford, J. P. A simple scoring weight for test items and its reliability. Psychometrika, 1941, 6, 367-374.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. Journal of Educational Measurement, 1971, 8, 9-14.

Hendrickson, G. F. An assessment of the effect of differential weighting options within items of a multiple-choice objective test using a Guttman-type weighting scheme. Unpublished doctoral dissertation, The Johns Hopkins University, 1970.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. Validity and reliability consequences of confidence weighting. Educational and Psychological Measurement, 1973, 33, 135-141.

Horn, J. L. A rationale and test for the number of factors in factor analysis. Psychometrika, 1965, 30, 179-186.

Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. Psychometrika, 1936, 1, 53-60.

Jacobs, S. S. Correlates of unwarranted confidence in responses to objective test items. Journal of Educational Measurement, 1971, 8, 15-20.

Johnson, P. O., & Jackson, R. W. Modern statistical methods: Descriptive and inductive. Chicago: Rand McNally & Co., 1959.

Kane, M. T., & Moloney, J. M. The effect of SSM grading on reliability when residual items have no discriminating power. Paper presented at the annual meeting of the National Council on Measurement in Education, April 1974.

Koehler, R. A. A comparison of the validities of conventional choice testing and various confidence marking procedures. Journal of Educational Measurement, 1971, 8, 297-303.

Koehler, R. A. Overconfidence on probabilistic tests. Journal of Educational Measurement, 1974, 11, 101-108.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Pugh, R. C., & Brunza, J. J. The contribution of selected personality traits and knowledge to response behavior on a probabilistic test. Paper presented at annual meeting of the American Educational Research Association, Chicago IL, April 1974.

Rippey, R. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-216.

Rippey, R. M. A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 1970, 7, 165-170.

Shuford, E.H., Albert, A., & Massengill, H.E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Slakter, M. J. Risk taking on objective examinations. American Educational

Research Journal, 1967, 4, 31-43.

Stanley, J. C., & Wang, M. D. Weighting test items and test-item options: An overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970, 30, 21-35.

Terwilliger, J. S. & Anderson, D. H. An empirical study of the effects of standardizing scores in the formation of linear composites. Journal of Educational Measurement, 6, 1969, 145-154.

Thompson, J. G., & Weiss, D. J. Criterion-related validity of adaptive testing strategies (Research Report 80-3). Minneapolis MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, June 1980.

Wang, M. W., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.

Wesman, A. G., & Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. Educational and Psychological Measurement, 1959, 19, 243-246.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6). Princeton NJ: Educational Testing Service, 1976.

Appendix:
Supplementary Tables

Table A
IRT Item Parameters for
Multiple-Choice Analogy Items

Item Number	a	b	c
310	.616	-.483	.20
273	.627	2.062	.20
275	.652	1.617	.21
286	.673	2.407	.09
327	.693	1.129	.22
399	.722	.446	.24
419	.750	2.413	.16
278	.770	2.002	.17
266	.815	1.690	.38
271	.828	1.266	.09
268	.844	1.036	.17
392	.865	-.360	.20
492	.914	-.145	.12
331	.930	1.352	.20
578	.946	.271	.20
405	.983	.739	.16
323	1.005	.828	.20
394	1.006	-.153	.20
277	1.041	1.930	.17
335	1.075	1.525	.20
575	1.098	.197	.25
560	1.132	-.007	.27
452	1.156	-.341	.30
493	1.172	.076	.26
576	1.211	.633	.20
415	1.234	1.183	.24
322	1.232	.960	.17
250	1.288	.513	.17
284	1.357	2.232	.24
339	1.608	1.818	.17
Mean	.975	.961	.20
SD	.244	.887	.06

Table B
Instructions Given Prior to Administration of Multiple-Choice Items with a Probabilistic Response Format

Screen 29891*
That completes the introductory information.

Type "GO" and press "RETURN" for the instructions for the first test.

Screen 29842*
This is a test of word knowledge. It is probably different from other tests you have taken, so it is important to read the instructions carefully to understand how to answer the questions.

Each question consists of a pair of words that have a specific relationship to each other, followed by four possible answers consisting of pairs of words. One of these four pairs of words has the same relationship as the first pair of words.

Type "GO" and press "RETURN" for an example.

Screen 29824*
For example:

Hot:Cold
1) Hard:Soft
2) Horse:Building
3) Mule:Horse
4) Yellow:Brown

Your job in this test is not to choose the correct answer (the pair of words that has the same relationship as the first pair of words) but to indicate your confidence that each of the four answers is the correct answer.

Type "GO" and press "RETURN" to continue the instructions.

Screen 29804*
You indicate your confidence by distributing 100 points among the four answers. The answer you think is the correct one should get the highest number of points, and the answer you feel is least likely to be the correct answer should get the lowest number of points.

The more certain you are that an answer is the correct one, the closer your response to that answer should be to 100. The more certain you are that an answer is NOT the correct one, the closer your response for that answer should be to 0.

-continued on the next page-

Table B, continued
Instructions Given Prior to Administration of Multiple-Choice Items with a Probabilistic Response Format

If you are completely certain that one of the answers is the correct answer, assign 100 to that answer and 0 to the other answers for that question. If you are completely uncertain as to which answer is correct, assign 25 to each of the four answers.

Type "GO" and press "RETURN" to continue.

Screen 29805*
The numbers you distribute among the four answers must sum to 99 or 100. However, you can distribute the 100 points in any way you like, as long as they reflect your certainty as to the "correctness" of each answer.

To answer a question, type the numbers you assign to each answer in a line in the order in which the answers appear in the question. Separate each number by a comma.

Type "GO" and press "RETURN" for an example.

Screen 29825*
Going back to the sample question:
Hot:Cold
1) Hard:Soft
2) House:Building
3) Mule:Horse
4) Yellow:Brown

Suppose a person responded with the following numbers:
? 80,0,0,20
This person was:

a) fairly sure, but not completely certain, that the first answer (Hard:Soft) had the same relationship as the pair of words in the question and thus was the correct answer.
b) completely certain that answers "2" and "3" were NOT the correct choice.
c) unsure about whether or not the fourth answer was the correct answer, but felt that it was closer to being an incorrect answer than the correct answer.

Note that 80 + 0 + 0 + 20 = 100.

Type "GO" and press "RETURN" to continue the instructions.

-continued on next page-

Table B, continued
Instructions Given Prior to Administration of Multiple-Choice Items with a Probabilistic Response Format

Screen 29826*
Let's look at this question once more:

Hot:Cold
1) Hard:Soft
2) House:Building
3) Mule:Horse
4) Yellow:Brown

Suppose a person responded with the following numbers:

? 33,0,33,33

This person was:
a) completely certain that the second answer was NOT the correct answer.
b) unsure as to which of the remaining answers was correct and felt that any of the remaining three answers were equally likely to be the correct answer.

Type "GO" and press "RETURN" to continue the instructions.

Screen 29827*
As you can see, there is an almost endless variety of combinations of numbers that you may use to state your confidence in the four possible answers. Use the entire range of numbers between 0 and 100 to express your confidence. Remember also that the numbers you assign to the four answers must sum to 99 or 100.

Please ask the proctor for help if you have any questions.

Type "GO" and press "RETURN" when you are ready to start the test.

*This line is for identification only and was not displayed.

Table C
Eigenvalues for the First Fifteen Principal Factors
of Real and Random Data for Each Scoring Method

Factor	QUAD Real	QUAD Random	SPHER Real	SPHER Random	TLOG Real	TLOG Random	PACA Real	PACA Random
1	6.38	1.01	6.67	1.00	5.65	1.02	8.16	1.04
2	1.23	.96	1.23	.96	1.36	.95	1.32	.96
3	.98	.93	.92	.94	1.21	.94	.96	.95
4	.93	.89	.91	.90	.97	.90	.80	.89
5	.84	.82	.81	.83	.89	.83	.71	.83
6	.74	.79	.72	.80	.81	.79	.65	.81
7	.69	.68	.71	.68	.73	.69	.60	.69
8	.67	.66	.66	.67	.72	.68	.56	.67
9	.63	.64	.61	.63	.63	.64	.55	.65
10	.57	.59	.55	.61	.58	.59	.50	.61
11	.47	.57	.47	.57	.49	.57	.45	.57
12	.44	.53	.43	.53	.47	.53	.42	.53
13	.41	.47	.42	.48	.42	.48	.38	.48
14	.40	.44	.39	.43	.42	.44	.36	.44
15	.38	.41	.35	.40	.39	.41	.30	.41

Distribution List

Navy

1 Liaison Scientist
Office of Naval Research
Branch Office, London
Box 39
FPO New York, NY 09510

1 Lt. Alexander Bory
Applied Psychology
Measurement Division
NAMRL
NAS Pensacola, FL 32508

1 Dr. Stanley Collyer
Office of Naval Technology
800 N. Quincy Street
Arlington, VA 22217

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 DR. PAT FEDERICO
Code P13
NPRDC
San Diego, CA 92152

1 Dr. Cathy Fernandes
Navy Personnel R&D Center
San Diego, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. William L. Maloy (02)
Chief of Naval Education and Training
Naval Air Station

Pensacola, FL 32508

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Cdr Ralph McCumber
Director, Research & Analysis Division
Navy Recruiting Command
4015 Wilson Boulevard
Arlington, VA 22203

1 Dr. George Moeller
Director, Behavioral Sciences Dept.
Naval Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 63409

1 Dr William Montague
NPRDC Code 13
San Diego, CA 92152

1 Bill Nordbrock
1032 Fairlawn Ave.
Libertyville, IL 60048

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

1 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychological Sciences Division
Code 442
Office of Naval Research
Arlington, VA 22217

6 Personnel & Training Research Group
Code 442PT
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
OP 115
Washington, DC 20350

1 LT Frank C. Petho, MSC, USN (Ph.D)
CNET (N-432)
NAS
Pensacola, FL 32508

1 Dr. Gary Poock
Operations Research Department
Code 55PK
Naval Postgraduate School
Monterey, CA 93940

1 Dr. Bernard Rimland (01C)
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

1 Dr. Worth Scanland
CNET (N-5)
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Frederick Steinheiser
CNO - OP115
Navy Annex
Arlington, VA 20370

1 Mr. Brad Sympson
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Frank Vicino
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Edward Wegman
Office of Naval Research (Code 411S&P)
800 North Quincy Street
Arlington, VA 22217

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Douglas Wetzel
Code 12
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152

Marine Corps

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Director, Office of Manpower Utilizatio
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Headquarters, U. S. Marine Corps
Code MPI-20
Washington, DC 20380

1 Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

1 Major Frank Yohannan, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Director, Training Research Lab
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joyce Shields
Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Hilda Wing
Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

1 Dr. Robert Wisher
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 AFHRL/LRS
Attn: Susan Ewing
WPAFB
WPAFB, OH 45433

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 U.S. Air Force Office of Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

1 Air University Library
AUL/LSE 76/443
Maxwell AFB, AL 36112

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Mr. Raymond E. Christal
AFHRL/MOE
Brooks AFB, TX 78235

1 Dr. Alfred R. Fregly
AFOSR/NL
Bolling AFB, DC 20332

1 Dr. Roger Pennell
Air Force Human Resources Laboratory
Lowry AFB, CO 80230

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

Department of Defense

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Jerry Lehnus
HQ MEPCOM
Attn: MEPCT-P
Fort Sheridan, IL 60037

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defens
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

Civilian Agencies

1 Dr. Helen J. Christup
Office of Personnel R&D
1900 E St., NW
Office of Personnel Management
Washington, DC 20015

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Chief, Psychological Reserch Branch
U. S. Coast Guard (G-P-1/2/TP42)
Washington, DC 20593

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Private Sector

1 Dr. James Algina
University of Florida
Gainesville, FL 326

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 Psychological Research Unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600
AUSTRALIA

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario, KOK
CANADA

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Mr. Arnold Bohrer
Section of Psychological Research
Caserne Petits Chateau
CRS
1000 Brussels
Belgium

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52243

1 Bundministerium der Verteidigung
-Referat P II 4-
Psychological Service
Postfach 1328
D-5300 Bonn 1
F. R. of Germany

1 Dr. Ernest R. Cadotte
307 Stokely
University of Tennessee
Knoxville, TN 37916

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

1 Dr. Kenneth B. Cross
Anacapa Sciences, Inc.
P.O. Drawer Q
Santa Barbara, CA 93102

1 Dr. Walter Cunningham
University of Miami
Department of Psychology
Gainesville, FL 32611

1 Dr. Dattpradad Divgi
Syracuse University
Department of Psychology
Syracuse, NE 33210

1 Dr. Fritz Drasgow
Department of Psychology
University of Illinois
603 E. Daniel St.
Champaign, IL 61820

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Dexter Fletcher
WICAT Research Institute
1875 S. State St.
Orem, UT 22333

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01002

1 Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3939 O'Hara Street
PITTSBURGH, PA 15260

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 DR. JAMES G. GREENO
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. Delwyn Harnisch
University of Illinois
242b Education
Urbana, IL 61801

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Dr. Jack Hunter
2122 Coolidge St.
Lansing, MI 48906

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Douglas H. Jones
Room T-255/21-T
Educational Testing Service
Princeton, NJ 08541

1 Professor John A. Keats
University of Newcastle
N. S. W. 2308
AUSTRALIA

1 Dr. Scott Kelso
Haskins Laboratories, Inc
270 Crown Street
New Haven, CT 06510

1 CDR Robert S. Kennedy
Canyon Research Group
1040 Woodcock Road
Suite 227
Orlando, FL 32803

1 Dr. William Koch
University of Texas-Austin
Measurement and Evaluation Center
Austin, TX 78703

1 Dr. Alan Lesgold
Learning R&D Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Mr. Phillip Livingston
Systems and Applied Sciences Corporatio
6811 Kenilworth Avenue
Riverdale, MD 20840

1 Dr. Robert Lockman
Center for Naval Analysis
200 North Beauregard St.
Alexandria, VA 22311

1 Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

1 Dr. James Lumsden
Department of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Mr. Robert McKinley
American College Testing Programs
P.O. Box 168
Iowa City, IA 52243

1

Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Dr. Robert Mislevy
711 Illinois Street
Geneva, IL 60134

1 Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Oklahoma City, OK 73069

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. James Olson
WICAT, Inc.
1875 South State Street
Orem, UT 84057

1 Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

1 Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Cirle, NW
Washington, DC 20036

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 Mr. L. Petrullo
3695 N. Nelson St.
ARLINGTON, VA 22207

1 Dr. Richard A. Pollak
Director, Special Projects
Minnesota Educational Computing
2520 Broadway Drive
St. Paul, MN

1 Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

1 Dr. Thomas Reynolds
University of Texas-Dallas
Marketing Department
P. O. Box 688
Richardson, TX 75080

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 Frank L. Schmidt
Department of Psychology
Bldg. GG
George Washington University
Washington, DC 20052

1 Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. William Sims
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Peter Stoloff
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. William Stout
University of Illinois
Department of Mathematics
Urbana, IL 61801

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN
THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Lab
252 Engineering Research Laboratory
Urbana, IL 61801

1 Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. V. R. R. Uppuluri
Union Carbide Corporation
Nuclear Division
P. O. Box Y
Oak Ridge, TN 37830

1 Dr. David Vale
Assessment Systems Corporation
2233 University Avenue
Suite 31
St. Paul, MN 55114

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Michael T. Waller
Department of Educational Psychology
University of Wisconsin--Milwaukee
Milwaukee, WI 53201

1 Dr. Brian Waters
HumRRO
300 North Washington
Alexandria, VA 22314

1 DR. GERSHON WELTMAN
PERCEPTRONICS INC.
6271 VARIEL AVE.
WOODLAND HILLS, CA 91367

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
Lawrence, KS 66045

1 Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90007

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

1 Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Previous Publications (Continued)

78-2. The Effects of Knowledge of Results and Test Difficulty on Ability Test Performance and Psychological Reactions to Testing. September 1978.
78-1. A Comparison of the Fairness of Adaptive and Conventional Testing Strategies. August 1978.
77-7. An Information Comparison of Conventional and Adaptive Tests in the Measurement of Classroom Achievement. October 1977.
77-6. An Adaptive Testing Strategy for Achievement Test Batteries. October 1977.
77-5. Calibration of an Item Pool for the Adaptive Measurement of Achievement. September 1977.
77-4. A Rapid Item-Search Procedure for Bayesian Adaptive Testing. May 1977.
77-3. Accuracy of Perceived Test-Item Difficulties. May 1977.
77-2. A Comparison of Information Functions of Multiple-Choice and Free-Response Vocabulary Items. April 1977.
77-1. Applications of Computerized Adaptive Testing. March 1977.
Final Report: Computerized Ability Testing, 1972-1975. April 1976.
76-5. Effects of Item Characteristics on Test Fairness. December 1976.
76-4. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. June 1976.
76-3. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. June 1976.
76-2. Effects of Time Limits on Test-Taking Behavior. April 1976.
76-1. Some Properties of a Bayesian Adaptive Ability Testing Strategy. March 1976.
75-6. A Simulation Study of Stradaptive Ability Testing. December 1975.
75-5. Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975.
75-4. A Study of Computer-Administered Stradaptive Ability Testing. October 1975.
75-3. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975.
75-2. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975.
75-1. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975.
74-5. Strategies of Adaptive Ability Measurement. December 1974.
74-4. Simulation Studies of Two-Stage Ability Testing. October 1974.
74-3. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974.
74-2. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974.
74-1. A Computer Software System for Adaptive Ability Measurement. January 1974.
73-4. An Empirical Study of Computer-Administered Two-Stage Ability Testing. October 1973.
73-3. The Stratified Adaptive Computerized Ability Test. September 1973.
73-2. Comparison of Four Empirical Item Scoring Procedures. August 1973.
73-1. Ability Measurement: Conventional or Adaptive? February 1973.

Copies of these reports are available, while supplies last, from:
Computerized Adaptive Testing Laboratory
N660 Elliott Hall
University of Minnesota
75 East River Road
Minneapolis MN 55455 U.S.A.

PREVIOUS PUBLICATIONS

Proceedings of the 1977 Computerized Adaptive Testing Conference. July 1978.

Proceedings of the 1979 Computerized Adaptive Testing Conference. September 1980.

Research Reports

83-2. Bias and Information of Bayesian Adaptive Testing. March 1983.

83-1. Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population. January 1983.

81-5. Dimensionality of Measured Achievement Over Time. December 1981.

81-4. Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries. November 1981.

81-3. A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing. September 1981.

Final Report: Computerized Adaptive Ability Testing. April 1981.

81-2. Effects of Immediate Feedback and Pacing of Item Presentation on Ability Test Performance and Psychological Reactions to Testing. February 1981.

81-1. Review of Test Theory and Methods. January 1981.

80-5. An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests. December 1980.

80-4. A Comparison of Adaptive, Sequential, and Conventional Testing Strategies for Mastery Decisions. November 1980.

80-3. Criterion-Related Validity of Adaptive Testing Strategies. June 1980.

80-2. Interactive Computer Administration of a Spatial Reasoning Test. April 1980.

Final Report: Computerized Adaptive Performance Evaluation. February 1980.

80-1. Effects of Immediate Knowledge of Results on Achievement Test Performance and Test Dimensionality. January 1980.

79-7. The Person Response Curve: Fit of Individuals to Item Characteristic Curve Models. December 1979.

79-6. Efficiency of an Adaptive Inter-Subtest Branching Strategy in the Measurement of Classroom Achievement. November 1979.

79-5. An Adaptive Testing Strategy for Mastery Decisions. September 1979.

79-4. Effect of Point-in-Time in Instruction on the Measurement of Achievement. August 1979.

79-3. Relationships among Achievement Level Estimates from Three Item Characteristic Curve Scoring Methods. April 1979.

Final Report: Bias-Free Computerized Testing. March 1979.

79-2. Effects of Computerized Adaptive Testing on Black and White Students. March 1979.

79-1. Computer Programs for Scoring Test Data with Item Characteristic Curve Models. February 1979.

78-5. An Item Bias Investigation of a Standardized Aptitude Test. December 1978.

78-4. A Construct Validation of Adaptive Achievement Testing. November 1978.

78-3. A Comparison of Levels and Dimensions of Performance in Black and White Groups on Tests of Vocabulary, Mathematics, and Spatial Ability. October 1978.

-continued inside-

END

FILMED

2-86

DTIC